Machine learning the apparent diffusion coefficient of Se(IV) in compacted bentonite

Xiaoqiong Shi[a], Junlei Tian[a], Jiacong Shen[a], Zhengye Feng[a], Jiaxing Feng[b], Tao Wu[*,a], Qingfeng Li[b]

[a] *Department of Engineering, Huzhou University, 313000 Huzhou, P. R. China*

[b] *Department of Science, Huzhou University, 313000 Huzhou, P. R. China*

Abstract

Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) algorithms were used to predict the apparent diffusion coefficient of Se(IV) in compacted bentonite. Seven instances of Se(IV) were measured using through-diffusion method. LightGBM (R^2 = 0.98 and RMSE = 0.025) exhibited superior predictive accuracy with a training dataset consisting of 956 instances and eight input features from Japan Atomic Energy Agency (JAEA-DDB). Shapley Additive Explanation and Partial Dependence Plots analyses revealed valuable insights into the diffusion mechanism of adsorbed anion obtained by evaluating the relationships between the apparent diffusion coefficient and the dependency of each input feature.

Keywords: Diffusion coefficient; Bentonite; Machine learning; Through-diffusion experiment.

1. Introduction

Bentonite is commonly regarded as engineering barriers in high-level radioactive

* Corresponding author. E-Mail address: twu@zjhu.edu.cn (Tao Wu)

waste repositories to hinder the release of radionuclides. The transportation of radionuclides in compacted bentonite is governed by diffusion-based mass transport, following Fick's law [1–3]. The retardation mechanism is controlled by the diffusion and sorption of radionuclides [4]. Apparent diffusion coefficient and effective diffusion coefficient are two crucial parameters for characterizing Fick diffusion [5]. The effective diffusion coefficient is calculated using Fick's first law, which involves analyzing radionuclide diffusion at steady-state stage. Since it takes a long time to reach the steady-state, a through-diffusion method is often employed for weak and non-sorbing radionuclides, such as ^{75}Se(IV) [6], HTO [3,7–9], Eu^{III}-$EDTA^{-}$ [10]. In contrast, the apparent diffusion coefficient measured at the transient-state stage considers the accessible porosity of bentonite in the accumulation term of Fick's second law for adsorbed radionuclides [4,5,11,12]. An In-diffusion method was commonly employed to measure the strongly sorbing radionuclides, such as ^{233}U(VI) [13], ^{237}Np(V) [14], and $^{134}Cs^{+}$ [15]. The relationship between the apparent diffusion coefficient (D_a) and the effective diffusion coefficient (D_e) can be expressed as follows [3]:

$$D_a = \frac{D_e}{\varepsilon + \rho_d \cdot K_d} = \frac{D_e}{\alpha}. \quad (1)$$

Where ε, ρ_d, K_d, and α are porosity, compacted dry density, distribution coefficient, and rock capacity factor.

Diffusion experiments are generally time-consuming and expensive for acquiring diffusion coefficients. The effective diffusion coefficient is often used in diffusion models, such as empirical equations and numerical models, enabling quick and cost-effective predictions. These models can quantitatively describe the dependence of the

effective diffusion coefficient of radionuclide species on mineral characteristics (mineral composition, porosity, cation exchange capacity, external surface area), solution properties (pH, Eh, mixed ions), and experimental conditions (compacted dry density, temperature, ionic strength) [6,11,12,16,17]. However, the predictive accuracy of these models remains unsatisfactory, which might be due to the approximation and assumptions during the modeling process. For example, the effective diffusion coefficient predicted by Archie's equation deviated from the experimental value by approximately 1–1.5 orders of magnitude mainly due to the approximation [18]. Numerical models, such as the integrated sorption and diffusion model [19], multi-porosity model [20,21], and pore-scale model [22,23], rarely provide a prediction accuracy index, which might due to the unsatisfied accuracy. Since the in-diffusion experiments for strongly sorbing radionuclides were more complex than the through-diffusion experiment [3,14,15], limited investigations have been conducted to measure the apparent diffusion coefficient. This parameter represents the coupling effect of radionuclide sorption and diffusion. As a result, there is a lack of information regarding its relationship with the aforementioned influencing factors.

Machine learning algorithms have been widely developed in regression prediction [20,24,25]. These algorithms utilized global analysis technologies, such as Shapley Additive Explanation, Individual Conditional Expectation, and Partial Dependence Plots analyses, to visually analyze the nonlinear relationship between ion diffusion coefficients and influencing factors [24,25]. Predictive accuracy is related to the dataset, including input features (influencing factors) and data size. Previous radionuclide

diffusion experiments primarily focused on specific experimental conditions, such as radionuclide species, compacted dry density, ionic strength, pH, temperature, and bentonite origin [3,7–10,17,26,27]. However, most investigations only considered a limited number of influencing factors. The data size decreased as the number of input features increased, posing challenges for the application of machine learning. This study examined the influence of input features and data size on predicting the apparent diffusion coefficient of Se(IV) in compacted bentonite. It utilized the through-diffusion method and two machine learning algorithms, namely Light Gradient Boosting Machine (LightGBM) and Random Forest (RF), based on a comprehensive diffusion dataset. Furthermore, the diffusion mechanism was investigated by analyzing the weight of influencing factors and the relationship between the apparent diffusion coefficient and each influencing factor.

2. Materials and Methods

2.1 Materials

Gaomiaozi (GMZ) bentonite, which was obtained from Beijing Research Institute of Uranium Geology, is originated from Gaomiaozi Mine in Xinghe Country (Inner Mongolia, China). It was converted to Ba-bentonite by mixing $BaCl_2$ and GMZ powder in a mass ratio of 5%. The detailed preparation procedure of Ba-bentonite was derived from a previous study [28]. They were compacted into Ø 25.6 × 12.6 mm blocks with the compacted dry density of 1300–1700 kg/m^3.

The stock solution of Se(IV) was prepared using SeO_2 (from Sinopharm Reagent)

powder dissolved in 0.1–0.5 mol/L NaCl solution. The initial concentration of Se(IV) in the diffusion experiment was $(12.5 \pm 1.0) \times 10^{-3}$ mol/L for Ba-bentonite experiments and $(17.9 \pm 0.8) \times 10^{-3}$ mol/L for GMZ experiments, respectively. All the solutions used in the experiment were prepared with ultrapure water from a Milli-Q system (Millipore, USA). The concentration of Se was determined by an Optima 7000DV inductively coupled plasma optical emission spectrometry (PerkinElmer, USA).

2.2 Diffusion experiments

Through-diffusion experiments were conducted under ambient condition at room temperature (22 ± 3°C). The bentonite blocks were completely saturated in 0.1–0.5 mol/L NaCl solution for five weeks. One side of diffusion cells (x = 0) was connected to a source reservoir filled with 200 mL of Se(IV) solution. The other side of the diffusion cell (x = L) was connected to a target reservoir with 10 mL of NaCl solution. Se(IV) diffused through the 12.6 mm thickness of a bentonite block and reached the target reservoir, which was regularly exchanged with a new 10 mL of NaCl solution without Se(IV) to keep the concentration gradient constant.

2.3 Diffusion data processing

Two sets of experimental data can be obtained: the first one consists of the accumulated mass (A_{cum}) of Se(IV) versus time, the second one consists of flux ($J(L,t)$) versus time [3,22]. The first one involves using the analytical solution of Fick's law to calculate the effective diffusion coefficient (D_e) and rock capacity factor (α), the equation is as follows [3]:

$$A_{cum} = S \cdot L \cdot C_0 \left(\frac{D_e \cdot t}{L^2} - \frac{\alpha}{6} - \frac{2 \cdot \alpha}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cdot \exp\left\{ - \frac{D_e \cdot n^2 \cdot \pi^2 \cdot t}{L^2 \cdot \alpha} \right\} \right), \quad (2)$$

where S, L, and C_0 were the cross section of bentonite block, the thickness of the block, and the initial concentration of Se(IV) in the source reservoir.

The second one was conducted to verify the effective diffusion coefficient and the rock capacity factor as follows:

$$J(L,t) = \frac{1}{S} \cdot \frac{\partial A_{cum}}{\partial t} . \quad (3)$$

2.4 Machine learning database description and analysis

The training dataset was sourced from the diffusion database system in Japan Atomic Energy Agency diffusion (JAEA-DDB) (1982–2009) and collected from relevant literatures on radionuclide diffusion in bentonite (2010–2023) [8,9,21,22,28–32]. Insufficient descriptions of influencing factors in JAEA-DDB and literatures led to a reduction in the number of instances as the influencing factors increased, significantly affecting the predictive performance of ML models. This study examined the number of instances and input features. Specifically, the input features numbered three, five, and eight, corresponding to a total of 850, 820, and 739 instances in JAEA-DDB (J3, J5, and J8). Additionally, 106 instances were collected from published literatures, resulting in total instances of 956, 926, and 845 for M3, M5, and M8, respectively. The dataset IM8 contained 956 instances that imputed M8 using the missForest imputation method [33]. The statistical information related to the number of input features and instances was summarized in **Table 1**. The input features included the compacted dry density, ion diffusion coefficient in water, ionic strength, rock

capacity factor, distribution coefficient, montmorillonite content, ionic charge, and temperature.

Table S1 in the supporting information provided a summary of the montmorillonite content for various types of bentonite. Since no literature montmorillonite content value of Kunipia-P was available, it was assumed to be equivalent to that of Na-montmorillonite [34]. The output variable was the apparent diffusion coefficient. Both ion diffusion coefficient in water and apparent diffusion coefficient were transformed into logarithmic form to enhance predictive accuracy.

To assess the predictive performance of machine learning models, the root mean square error (RMSE) and coefficient of determination (R^2) were utilized. These metrics can be calculated using the following equations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left(\mathrm{Log}D_{a,i}^{\ exp} - \mathrm{Log}D_{a,i}^{\ pred}\right)^2}, \tag{4}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{\mathrm{N}}\left(\mathrm{Log}D_{a,i}^{\ exp} - \mathrm{Log}D_{a,i}^{\ pred}\right)^2}{\sum_{i=1}^{\mathrm{N}}\left(\mathrm{Log}D_{a,i}^{\ exp} - \mathrm{Log}D_{a,ave}^{\ exp}\right)^2}, \tag{5}$$

with the number of samples (N), the logarithms of the experimental and predicted apparent diffusion coefficient ($\mathrm{Log}D_{a,i}^{\ exp}$ and $\mathrm{Log}D_{a,i}^{\ pred}$), and the average experimental apparent diffusion coefficient ($\mathrm{Log}D_{a,ave}^{\ exp}$). Models demonstrating great accuracy and performance are characterized by low RMSE and high R^2.

151

Table 1 Descriptive statistics of the dataset for each model.

		Input								Output
	Model	Instance number	ρ_d (kg/m^3)	LogD_w (–)	I (mol/L)	α (–)	K_d (m^3/kg)	m (–)	z (–)	T (°C)	LogD_a (–)
Mean	J3	850	1371	–8.78	0.24	–	–	–	–	–	–10.49
	M3	956	1368	–8.78	0.25	–	–	–	–	–	–10.48
	J5	820	1366	–8.79	0.24	124	0.08	–	–	–	–10.52
	M5	926	1363	–8.79	0.25	115	0.07	–	–	–	–10.50
	J8	739	1298	–8.78	0.18	137	0.09	0.85	0.32	29.31	–10.50
	M8	845	1304	–8.78	0.19	126	0.08	0.85	0.21	28.51	–10.49
	IM8	956	1368	–8.78	0.25	121	0.08	0.84	0.14	27.89	–10.48
Std	J3	850	446	0.20	0.69	–	–	–	–	–	0.93
	M3	956	439	0.19	0.66	–	–	–	–	–	0.91
	J5	820	446	0.21	0.70	1310	0.67	–	–	–	0.93
	M5	926	439	0.20	0.66	1235	0.63	–	–	–	0.90
	J8	739	389	0.21	0.43	1380	0.70	0.20	1.31	16.32	0.93
	M8	845	388	0.20	0.42	1292	0.66	0.20	1.29	15.54	0.90
	IM8	956	439	0.19	0.66	1239	0.63	0.19	1.29	14.74	0.91
Min	J3	850	400	–9.30	0	–	–	–	–	–	–15.55
	M3	956	400	–9.30	0	–	–	–	–	–	–15.55
	J5	820	400	–9.30	0	0.03	-3.8×10^{-4}	–	–	–	–15.55
	M5	926	400	–9.30	0	0.02	-3.8×10^{-4}	–	–	–	–15.55
	J8	739	400	–9.30	0	0.05	-3.8×10^{-4}	0	–2	5	–15.55
	M8	845	400	–9.30	0	0.02	-3.8×10^{-4}	0	–2	5	–15.55
	IM8	956	400	–9.30	0	0.01	-3.8×10^{-4}	0	-2	5	–15.55

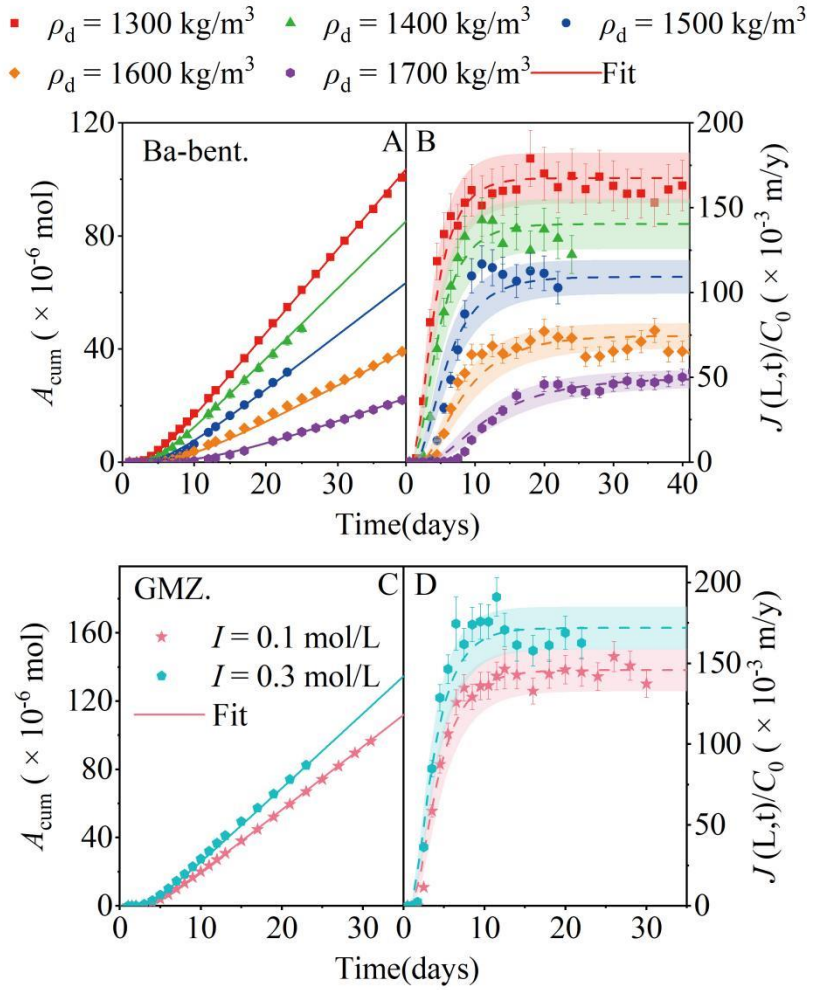
Max	J3	850	2730	–8.24	5	–	–	–	–	–	–8.97
	M3	956	2730	–8.24	5	–	–	–	–	–	–8.97
	J5	820	2730	–8.24	5	34877	17.40	–	–	–	–8.97
	M5	926	2730	–8.24	5	34877	17.40	–	–	–	–8.97
	J8	739	2330	–8.24	5	34877	17.40	1	5	90	–8.97
	M8	845	2330	–8.24	5	34877	17.40	1	5	90	–8.97
	IM8	956	2730	–8.24	5	34877	17.4	1	5	90	–8.97
Skw	J3	850	0.59	–0.21	6	–	–	–	–	–	–1.53
	M3	956	0.52	–0.22	6.10	–	–	–	–	–	–1.53
	J5	820	0.57	–0.17	5.92	23.41	22.01	–	–	–	–1.55
	M5	926	0.50	–0.18	6.02	24.79	22.89	–	–	–	–1.55
	J8	739	0.22	–0.14	7.92	22.24	20.91	–2.27	0.47	1.32	–1.65
	M8	845	0.13	–0.15	7.48	23.69	21.88	–2.10	0.60	1.46	–1.64
	IM8	956	0.52	–0.22	6.10	23.97	22.51	–2.01	0.63	1.62	–1.53

Std = Standard Deviation; Skw = Skewness

3. Results and discussion

3.1. Measurement of the apparent diffusion coefficient by through-diffusion experiments

The through-diffusion method has been extensively used in the study of the diffusion of anionic radionuclides, attributed to the high diffusivity resulting from the anionic exclusion effect [9,10,22]. ^{79}Se is a crucial radionuclide in repository safety assessment. The primary form of Se(IV) was $HSeO_3^-$ at pH ranging from 4 to 7 [35]. **Fig. 1** illustrates the breakthrough curves of $HSeO_3^-$ in Ba-bentonite and GMZ bentonite. The accumulated mass of $HSeO_3^-$ exhibited a linear increase during the steady-state stage, which is related to the effective diffusion coefficient. In contrast, the flux showed a significant increase during the transient-state stage, which is linked to the rock capacity factor. The time taken to reach the steady-state stage is related to the compacted dry density and increase from about 10 days at the compacted dry density of 1300 kg/m^3 to about 20 days at the compacted dry density of 1700 kg/m^3.



168

169 **Fig. 1** The accumulated mass (A_{cum}) and flux ($J(L,t)$) of $HSeO_3^-$ in compacted bentonite

170 as a function of time. (A and B) Ba-bentonite, I = 0.5mol/L, pH = 3.1 ± 0.1, T = 22 ±

171 3°C. (C and D) GMZ, ρ_d = 1300 kg/m^3, pH = 5.6 ± 0.1, T = 22 ± 3°C.

172 **Table 2** lists the diffusion parameters of $HSeO_3^-$ in compacted Ba-bentonite and

173 GMZ bentonite. The apparent diffusion coefficient of Ba-bentonite decreased from 6.8

174 × 10^{-11} m^2/s to 2.2 × 10^{-11} m^2/s as the compacted dry density increased from 1300 kg/m^3

175 to 1700 kg/m^3. In addition, the apparent diffusion coefficient of GMZ bentonite

176 increased from 5.9 × 10^{-11} m^2/s to 6.7 × 10^{-11} m^2/s as the ionic strength increased from

0.1 mol/L to 0.3 mol/L. It was slightly lower than that reported in synthetic pore water (D_a = 7.8 × 10^{-11} m^2/s) in a previous study [36], which could attribute to the higher pH and more complexity of the ions in the synthetic pore water.

Table 2 Summary of diffusion parameters in Ba-bentonite and GMZ bentonite.

Bentonite	C_0 (× 10^{-3} mol/L)	I (mol/L)	ρ_d (kg/m^3)	D_e (× 10^{-11} m^2/s)	α (−)	K_d (× 10^{-4} m^3/kg)	D_a (× 10^{-11} m^2/s)
Ba-bent.	12.5 ± 1.0	0.5	1300	6.8 ± 0.6	1.00 ± 0.08	3.8 ± 0.3	6.8 ± 0.8
		0.5	1400	5.7 ± 0.6	0.96 ± 0.08	3.5 ± 0.4	5.9 ± 0.8
		0.5	1500	4.4 ± 0.4	0.96 ± 0.08	3.5 ± 0.6	4.6 ± 0.6
		0.5	1600	3.0 ± 0.3	0.92 ± 0.08	3.3 ± 0.2	3.3 ± 0.4
		0.5	1700	2.0 ± 0.2	0.90 ± 0.08	3.2 ± 0.4	2.2 ± 0.3
GMZ	17.9 ± 0.8	0.1	1300	5.6 ± 0.5	0.95 ± 0.08	3.2 ± 0.3	5.9 ± 0.7
		0.3	1300	6.6 ± 0.5	0.98 ± 0.08	3.4 ± 0.3	6.7 ± 0.8

3.2. Prediction of machine learning algorithms

Both LightGBM and RF are popular and powerful tree-based machine learning algorithms employed in regression analysis. They have been applied in various diffusion studies, such as predicting the effective diffusion coefficient of Re(VII) in compacted bentonite [20] and chloride diffusion coefficient in cements [25]. In this study, they were employed to predict the apparent diffusion coefficient of $HSeO_3^-$ in compacted bentonite. Hyperparameters refer to a set of values that are specified before training the machine learning. Optimal performance can be achieved through successful hyperparameter optimization. **Table 3** lists the optimized hyperparameters for LightGBM and RF algorithms.

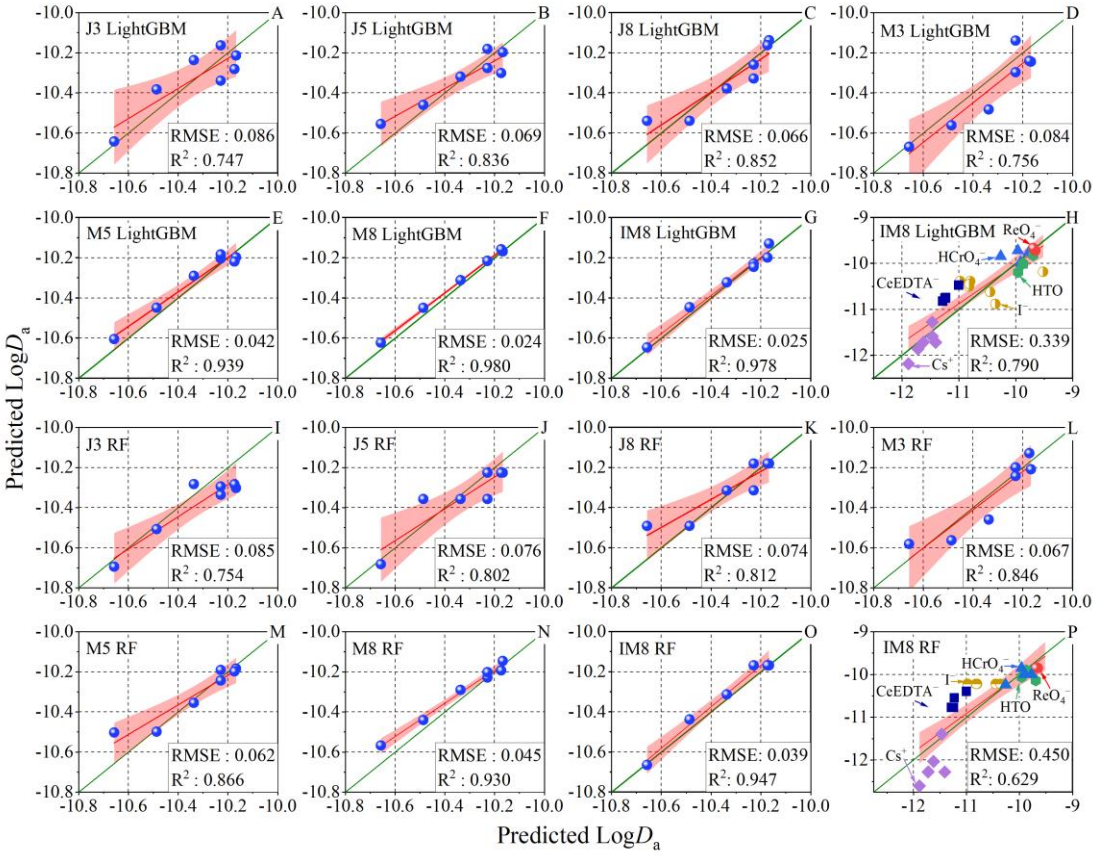
Table 3 The optimal parameters for Light Gradient Boosting Machine and Random Forest algorithms.

Algorithms	Parameters	J3	J5	J8	M3	M5	M8	IM8
LightGBM	Num_leaves	30	30	30	31	31	30	30
	Min_data_in_leaf	17	11	27	53	29	20	3
	Max_depth	3	6	3	–1	11	–1	5
	Learning_rate	0.30	0.30	0.25	0.03	0.01	0.10	0.26
	Num_boost_round	10000	10000	10000	10000	10000	10000	10000
	Feature_fraction	0.60	0.48	0.37	1	1	1	0.39
RF	Max_depth	None	8	6	None	None	15	7
	Min_samples_split	2	5	3	2	2	3	12
	Max_features	Auto	Auto	Auto	Auto	Auto	Auto	0.42
	N_estimators	10	10	3	14	10	12	3

The predictive results of the apparent diffusion coefficient using LightGBM and RF for different training sets are shown in **Fig. 2** (J3, J5, J8, M3, M5, M8, and IM8). **Figs. 2A**–**2G** show the prediction results of LightGBM and **Figs. 2I**–**2O** show the prediction results of RF. **Figs. 2H** and **2P** represent the application of IM8 in predicting the D_a values of ReO_4^-, $HCrO_4^-$, I^-, $CeEDTA^-$, HTO, and Cs^+ in compacted bentonite [37–40]. These species are selected to surrogate the monovalent radionuclide anionic species, neutral molecules, and monovalent cations. Notably, the M-model exhibited lower RMSE and higher R^2 compared to the J-model, indicating that a larger number of instances enhanced predictive performance. The RMSE ranked in the following order: J3 > J5 > J8 and M3 > M5 > M8, while the R^2 ranked in the opposite order of RMSE: J3 < J5 < J8 and M3 < M5 < M8. The predictive accuracy of IM8 was similar to M8 for the LightGBM algorithm, while IM8 exhibited higher predictive accuracy than M8 for the RF algorithm. These observations indicate that an increased number of input features enhanced predictive performance [41].

For J3 and M3, both LightGBM and RF produced relatively low prediction

accuracy with R^2 below 0.8, indicating that three input features were insufficient for accurate predictions of the apparent diffusion coefficient for $HSeO_3^-$ in compacted bentonite. However, when the number of input features increased to five, the predictive accuracy of M5 using LightGBM significantly increased with R^2 of 0.939 and RMSE of 0.042. The IM8 demonstrated superior predictive accuracy for both LightGBM (R^2 = 0.98 and RMSE = 0.025) and RF (R^2 = 0.95 and RMSE = 0.039). The application of IM8 in predicting the D_a values of radionuclides also exhibited a good prediction performance, with R^2 of 0.79 and RMSE of 0.34 for LightGBM and R^2 of 0.63 and RMSE of 0.45 for RF (**Figs. 2H** and **2P**). It indicates that LightGBM has better generalization ability than RF. LightGBM utilizes gradient-based one-sided sampling, exclusive feature bundling, and histogram-based algorithm techniques, effectively preventing overfitting, accelerating training speed, and reducing memory consumption, thereby leading to improved predictive accuracy.



223

224 **Fig. 2** The apparent diffusion coefficient of $HSeO_3^-$ in compacted bentonite was
225 predicted using Light Gradient Boosting Machine and Random Forest for various
226 numbers of input features and instances. The experimental D_a values of ReO_4^-, $HCrO_4^-$,
227 I^-, $CeEDTA^-$, HTO, and Cs^+ in Figs. 2H and 2P were from [37–40].

228 ### 3.3. Spearman and Shapley Additive Explanation analyses

229 **Fig. 3** shows the correlation of each feature with D_a in the training dataset using
230 Spearman analysis and the weight of input features for IM8 using Shapley Additive
231 Explanation (SHAP) analysis. Models of J3–J8 and M3–IM8 show similar Spearman's
232 correlation coefficients, indicating that the number of input features and instances had
233 insignificant on the non-linear relationship (**Fig. 3A**). Four parameters (compacted dry
234 density, rock capacity factor, distribution coefficient, and ionic charge) had negatively

correlated with the apparent diffusion coefficient, three parameters (ion diffusion coefficient in water, montmorillonite content, and temperature) exhibited positive correlation, and ionic strength had insignificant impact. The rock capacity factor, distribution coefficient, compacted dry density, and ionic charge had high absolute values of Spearman's correlation coefficients, indicating their high correlations with the apparent diffusion coefficient.

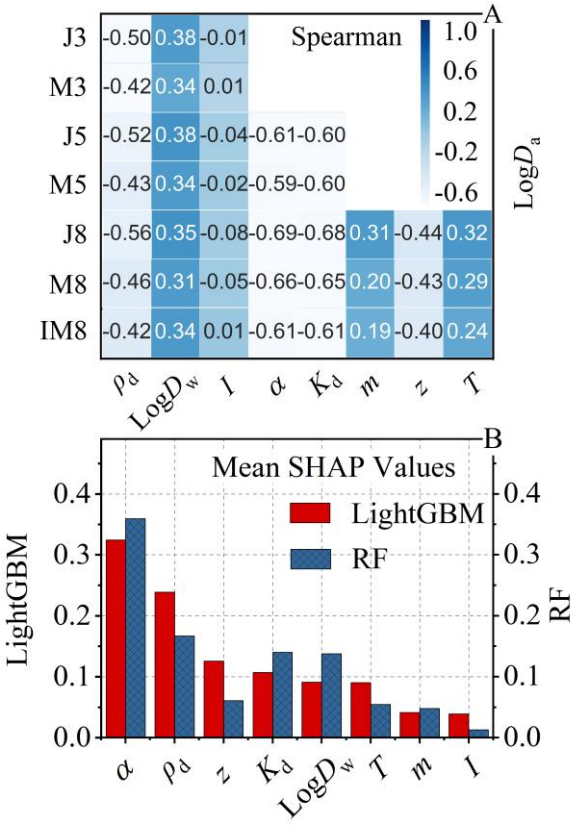


Fig. 3 (A) The correlation analysis of the training dataset and (B) the contribution of input features to the predicted apparent diffusion coefficient.

Shapley Additive Explanations (SHAP) method is commonly utilized for analyzing the weight or importance of input features on prediction. The results of SHAP in LightGBM and RF demonstrated an approximately similar ranking order (**Fig. 3B**). Specifically, LightGBM ranked the ion diffusion coefficient in water the fifth, whereas

RF ranked it the fourth. The top two input features for predicting apparent diffusion coefficient were the rock capacity factor and compacted dry density, contributing to a total of 53.3% for LightGBM and 53.7% for RF compared to all features in IM8.

The montmorillonite content and ionic strength are two important parameters that has been extensively investigated. Since they have a significant influence on the effective diffusion coefficient of anionic radionuclides [8,22,42]. SHAP analysis demonstrates that the montmorillonite content and ionic strength made an insignificant contribution to the prediction (**Fig. 3B**). The tendency of ionic strength is consistent with [9], who reported that the insignificant influence on the apparent diffusion coefficient of Cs^+ and Na^+ was attributed to the coupled effects of diffusion and sorption. Further studies are needed to verify these results, as the weight of input features is influenced by different types and numbers of instances, and machine learning algorithms.

3.4. Partial Dependence Plots analysis

Partial Dependence Plots (PDP) can visualize the relationship between each input feature and the predicted apparent diffusion coefficient. **Fig. 4** shows the contribution of each input feature to the prediction for IM8 using PDP analysis. The gray column shows the distribution of data points of an input feature. The PDP value of LightGBM is represented by solid lines, while the dashed line represents the PDP value of RF. A flat curve suggests that the feature has an insignificant influence on the predicted outcome. In contrast, a steep curve indicates a stronger relationship. Both LightGBM

and RF demonstrate similar trends for the dependency of predicted apparent diffusion coefficient on each input feature.

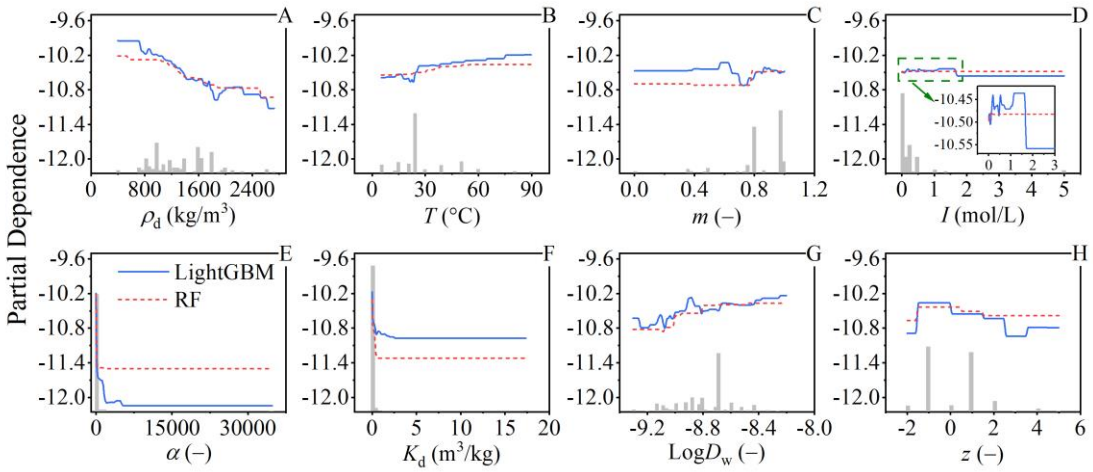


Fig. 4 Partial Dependence Plots (PDP) analysis based on the Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) applied to the prediction of the apparent diffusion coefficient: (A) compacted dry density, (B) temperature, (C) montmorillonite content, (D) ionic strength, (E) rock capacity factor, (F) distribution coefficient, (G) ion diffusion coefficient in water, and (H) ionic charge.

The analysis of the relationship between the predicted apparent diffusion coefficient and the input features related to experimental conditions was consistent with published experimental results [3,13,26]. Specifically, it showed a negative correlation with the compacted dry density (**Fig. 4A**). It can be explained that the available pores for radionuclide diffusion decreased with increasing compaction. At the compacted dry density above 1600 kg/m^3, the decreasing tendency changed slowly, attributed to the reduction of interlayer pores to one water layer [21,42,43].

The apparent diffusion coefficient exhibited a positive correlation with temperature (**Fig. 4B**). In a repository, the temperature of clay closed to the waste containers could exceed 100°C [44]. The positive relationship between the effective diffusion coefficient and temperature was reported for HTO, $^{36}Cl^-$, and ReO_4^-, which followed Arrhenius equation [3,7,26]. **Figs. 4C** and **4D** show that the apparent diffusion coefficient exhibited insignificant correlations with montmorillonite content and ionic strength. Slight negative and positive impacts were observed for montmorillonite content ranged from 0.5 to 0.8 and ionic strength below 1.0, respectively, which agree with their correlations with the effective diffusion coefficient [20]. Plenty of studies have shown that as the ionic strength increases, the effective diffusion coefficient of the radionuclide also increases [8,22]. This tendency was explained by the fact that the thickness of electrical double layer decreased in high salinity solution [9,22]. In general, the consistency between PDP and experimental results implies that PDP analysis can provide insights into the diffusion law and diffusion mechanism of radionuclides.

The input features, namely the rock capacity factor, distribution coefficient, ion diffusion coefficient in water, and ionic charge, are the parameters related to the properties of radionuclides. Their dependencies on the apparent diffusion coefficient were unclear due to the coupled effect of solid, liquid, and radionuclides. **Fig. 4E** indicates that the rock capacity factor had a negative impact on the prediction. According to Eq. (3), the apparent diffusion coefficient is inversely proportional to the rock capacity factor. The negative relationship between apparent diffusion coefficient and distribution coefficient can be attributed to the linear correlation of the rock

capacity factor and distribution coefficient (**Fig. 4F**). The predicted apparent diffusion coefficient increased with increasing ion diffusion coefficient in water (**Fig. 4G**), indicating that radionuclide diffuse quickly in both liquid and solid. RF algorithm shows that the ionic charge had insignificant on apparent diffusion coefficient, whereas LightGBM algorithm demonstrates that it decreased as the ionic charge reduced from 0 to +3 (**Fig. 4H**).

4. Conclusions

The effect of input features and instances on the prediction of apparent diffusion coefficient was conducted using Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) algorithms. $HSeO_3^-$ (as a surrogate to $^{79}HSeO_3^-$) diffusion experiment in compacted bentonite was conducted using a through-diffusion method to testify the predictive performance. Increasing the number of input features resulted in a decrease of instances. LightGBM (R^2 = 0.98 and RMSE = 0.025) and RF (R^2 = 0.95 and RMSE = 0.039) exhibited superior predictive accuracy for the training set of 956 instances and eight input features, which were the compacted dry density, ion diffusion coefficient in water, ionic strength, rock capacity factor, distribution coefficient, montmorillonite content, ionic charge, and temperature.

Shapley Additive Explanations exhibited that the top two input features for predicting apparent diffusion coefficient were the rock capacity factor and compacted dry density. Partial Dependence Plots indicated that the dependency of apparent

diffusion coefficient on the rock capacity factor, compacted dry density, and distribution coefficient was negative, whereas the positive relationship between the apparent diffusion coefficient and the ion diffusion coefficient in water was observed. This study presented a method for predicting the apparent diffusion coefficient, quantifying the influencing factors, and understanding the effect of each input features on the apparent diffusion coefficient. The insightful information on the diffusion mechanism is beneficial for the safety assessment of repositories.

Acknowledgements

This work was support from the key program of national natural science foundation of China (12335008), Huzhou science and technology planning project (2021GZ60), and Scientific Research and Innovation Project for postgraduate by School of Engineering, Huzhou University.

References

1. Benning JL, Barnes DL (2009) Comparison of modeling methods for the determination of effective porosities and diffusion coefficients in through-diffusion tests. Water Resour Res 45(9): W09419. https://doi.org/10.1029/2008wr007236
2. Shackelford CD, Moore SM (2013) Fickian diffusion of radionuclides for engineered containment barriers: diffusion coefficients, porosities, and complicating issues. Eng Geol 152(1):133−147.

https://doi.org/10.1016/j.enggeo.2012.10.014

3. Van Loon LR, Soler JM (2003) Diffusion of HTO, $^{36}Cl^{-}$, $^{125}I^{-}$ and $^{22}Na^{+}$ in opalinus clay: Effect of confining pressure, sample orientation, sample depth and temperature. Paul Scherrer Institut (PSI):1015−2636

4. Liu Z, Emami-Meybodi H (2022) Apparent diffusion coefficient for adsorption-controlled gas transport in nanoporous media. Chem Eng J 450:138105. https://doi.org/10.1016/j.cej.2022.138105

5. Cormenzana JL, García Gutiérrez M, Missana T, Junghanns Á (2003) Simultaneous estimation of effective and apparent diffusion coefficients in compacted bentonite. J Contam Hydrol 61(1−4):63−72. https://doi.org/10.1016/s0169-7722(02)00113-4

6. Yang X, Ge X, He J, Wang C, Qi L, Wang X, Liu C (2018) Effects of mineral compositions on matrix diffusion and sorption of $^{75}Se(IV)$ in Granite. Environ Sci Technol 52(3):1320−1329. https://doi.org/10.1021/acs.est.7b05795

7. Bestel M, Glaus MA, Frick S, Gimmi T, Juranyi F, Van Loon LR, Diamond LW (2018) Combined tracer through-diffusion of HTO and ^{22}Na through Na-montmorillonite with different bulk dry densities. Appl Geochem 93:158−166. https://doi.org/10.1016/j.apgeochem.2018.04.008

8. Fukatsu Y, Yotsuji K, Ohkubo T, Tachi Y (2021) Diffusion of tritiated water, $^{137}Cs^{+}$, and $^{125}I^{-}$ in compacted Ca-montmorillonite: Experimental and modeling approaches. Appl Clay Sci 211:106176. https://doi.org/10.1016/j.clay.2021.106176

9. Tachi Y, Yotsuji K (2014) Diffusion and sorption of Cs^+, Na^+, I^- and HTO in compacted sodium montmorillonite as a function of porewater salinity: integrated sorption and diffusion model. Geochim Cosmochim Acta 132:75−93. https://doi.org/10.1016/j.gca.2014.02.004

10. Descostes M, Pointeau I, Radwan J, Poonoosamy J, Lacour JL, Menut D, Vercouter T, Dagnelie RVH (2017) Adsorption and retarded diffusion of Eu^{III}-$EDTA^-$ through hard clay rock. J Contam Hydrol 544:125−132. https://doi.org/10.1016/j.jhydrol.2016.11.014

11. Cho WJ, Oscarson DW, Hahn PS (1993) The measurement of apparent diffusion coefficients in compacted clays: an assessment of methods. Appl Clay Sci 8(4):283−294. https://doi.org/10.1016/0169-1317(93)90009-P

12. Gooddy DC, Kinniburgh DG, Barker JA (2007) A rapid method for determining apparent diffusion coefficients in chalk and other consolidated porous media. J Hydrol 343(1−2):97−103. https://doi.org/10.1016/j.jhydrol.2007.06.014

13. Joseph C, Van Loon LR, Jakob A, Steudtner R, Schmeide K, Sachs S, Bernhard G (2013) Diffusion of U(VI) in opalinus clay: influence of temperature and humic acid. Geochim Cosmochim Acta 109:74−89. https://doi.org/10.1016/j.gca.2013.01.027

14. Wu T, Amayri S, Drebert J, Van Loon LR, Reich T (2009) Neptunium(V) sorption and diffusion in opalinus clay. Environ Sci Technol 43:6567−6571. https://doi.org/10.1021/es9008568

15. Van Loon LR, Müller W (2014) A modified version of the combined in-

394 diffusion/abrasive peeling technique for measuring diffusion of strongly sorbing
395 radionuclides in argillaceous rocks: A test study on the diffusion of caesium in
396 Opalinus Clay. Appl Radiat Isot 90:197−202.
397 https://doi.org/10.1016/j.apradiso.2014.04.009
398 16. Cui LY, Masum SA, Ye WM, Thomas HR (2021) Investigation on gas migration
399 behaviours in saturated compacted bentonite under rigid boundary conditions.
400 Acta Geotech 17(6):2517−2531. https://doi.org/10.1007/s11440-021-01424-1
401 17. Kozaki T, Sawaguchi T, Fujishima A, Sato S (2010) Effect of exchangeable
402 cations on apparent diffusion of Ca^{2+} ions in Na- and Ca-montmorillonite
403 mixtures. Phys Chem Earth 35(6−8):254−258.
404 https://doi.org/10.1016/j.pce.2010.04.006
405 18. Van Loon LR, Mibus J (2015) A modified version of Archie's law to estimate
406 effective diffusion coefficients of radionuclides in argillaceous rocks and its
407 application in safety analysis studies. Appl Geochem 59:85−94.
408 https://doi.org/10.1016/j.apgeochem.2015.04.002
409 19. Ochs M, Lothenbach B, Wanner H, Sato H, Yui M (2001) An integrated
410 sorption-diffusion model for the calculation of consistent distribution and
411 diffusion coefficients in compacted bentonite. J Contam Hydrol
412 47(2−4):283−296. https://doi.org/10.1016/S0169-7722(00)00157-1
413 20. Feng Z, Gao Z, Wang Y, Wu T, Li Q (2023) Application of machine learning to
414 study the effective diffusion coefficient of Re(VII) in compacted bentonite.
415 Appl Clay Sci 243:107076. https://doi.org/10.1016/j.clay.2023.107076

21. Wu T, Wang Z, Tong Y, Wang Y, Van Loon LR (2018) Investigation of Re(VII) diffusion in bentonite by through-diffusion and modeling techniques. Appl Clay Sci 166:223−229. https://doi.org/10.1016/j.clay.2018.08.023

22. Wu T, Yang Y, Wang Z, Shen Q, Tong Y, Wang M (2020) Anion diffusion in compacted clays by pore-scale simulation and experiments. Water Resour Res 56(11):e2019WR027037. https://doi.org/10.1029/2019wr027037

23. Yang Y, Churakov SV, Patel RA, Prasianakis N, Deissmann G, Bosbach D, Poonoosamy J (2024) Pore-Scale Modeling of Water and Ion Diffusion in Partially Saturated Clays. Water Resour Res 60(1): e2023WR035595. https://doi.org/10.1029/2023wr035595

24. Mohammadi Golafshani E, Kashani A, Kim T, Arashpour M (2022) Concrete chloride diffusion modelling using marine creatures-based metaheuristic artificial intelligence. J Cleaner Prod 374:134021. https://doi.org/10.1016/j.jclepro.2022.134021

25. Tran VQ (2022) Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials. Constr Build Mater 328:127103. https://doi.org/10.1016/j.conbuildmat.2022.127103

26. Wu T, Wang Z, Li Q, Pan G, Li J, Van Loon LR (2016) Re(VII) diffusion in bentonite: effect of organic compounds, pH and temperature. Appl Clay Sci 127−128:10−16. https://doi.org/10.1016/j.clay.2016.03.039

27. Tian W, Li C, Liu X, Wang L, Zheng Z, Wang X, Liu C (2012) The effect of ionic strength on the diffusion of ^{125}I in Gaomiaozi bentonite. J Radioanal Nucl

Chem 295(2):1423−1430. https://doi.org/10.1007/s10967-012-2284-y

28. Wu T, Feng Z, Geng Z, Xu M, Shen Q (2023) Restriction of Re(VII) and Se(IV) diffusion by barite precipitation in compacted bentonite. Appl Clay Sci 232:106803. https://doi.org/10.1016/j.clay.2022.106803

29. Geng Z, Feng Z, Li H, Wang Y, Wu T (2022) Porosity investigation of compacted bentonite using through-diffusion method and multi-porosity model. Appl Geochem 146:105480. https://doi.org/10.1016/j.apgeochem.2022.105480

30. Glaus MA, Frick S, Rossé R, Van Loon LR (2010) Comparative study of tracer diffusion of HTO, $^{22}Na^{+}$ and $^{36}Cl^{-}$ in compacted kaolinite, illite and montmorillonite. Geochim Cosmochim Acta 74(7):1999−2010. https://doi.org/10.1016/j.gca.2010.01.010

31. Tachi Y, Nakazawa T, Ochs M, Yotsuji K, Suyama T, Seida Y, Yamada N, Yui M (2010) Diffusion and sorption of neptunium(V) in compacted montmorillonite: effects of carbonate and salinity. Radiochim Acta 98(9−11):711−718. https://doi.org/10.1524/ract.2010.1772

32. Wu T, Wang Z, Wang H, Zhang Z, Van Loon LR (2017) Salt effects on Re(VII) and Se(IV) diffusion in bentonite. Appl Clay Sci 141:104−110. https://doi.org/10.1016/j.clay.2017.02.021

33. Joel LO, Doorsamy W, Paul BS (2024) On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets. ArXiv. https://doi.org/10.48550/arXiv.2403.14687

34. González Sánchez F, Van Loon LR, Gimmi T, Jakob A, Glaus MA, Diamond

460 LW (2008) Self-diffusion of water and its dependence on temperature and ionic
461 strength in highly compacted montmorillonite, illite and kaolinite. Appl
462 Geochem 23(12):3840−3851.
463 https://doi.org/10.1016/j.apgeochem.2008.08.008
464 35. Shi K, Ye Y, Guo N, Guo Z, Wu W (2013) Evaluation of Se(IV) removal from
465 aqueous solution by GMZ Na-bentonite: batch experiment and modeling studies.
466 J Radioanal Nucl Chem 299(1):583–589. https://doi.org/10.1007/s10967-013-
467 2807-1
468 36. Wu T, Wang H, Zheng Q, Zhao YL, Van Loon LR (2014) Diffusion behavior of
469 Se(IV) and Re(VII) in GMZ bentonite. Appl Clay Sci 101:136−140.
470 https://doi.org/10.1016/j.clay.2014.07.028
471 37. Tachi Y, Yotsuji K, Seida Y, Yui M (2009) Diffusion of cesium and iodine in
472 compacted sodium montmorillonite under different saline conditions. MRS
473 Proceedings 1193:545−552. http://dx.doi.org/10.1557/PROC-1193-545
474 38. Feng Z, Tian J, Wu T, Wei G, Li Z, Shi X, Wang Y, Li Q (2024) Unveiling the
475 Re, Cr, and I diffusion in saturated compacted bentonite using machine-learning
476 methods. Nucl Sci Tech. Accepted
477 39. Joseph C, Mibus J, Trepte P, Müller C, Brendler V, Park DM, Jiao Y, Kersting
478 AB, Zavarin M (2017) Long-term diffusion of U(VI) in bentonite: Dependence
479 on density. Sci Total Environ 575:207−218.
480 https://doi.org/10.1016/j.scitotenv.2016.10.005
481 40. Wu T, Hong Y, Shao D, Zhao J, Feng Z (2023) Experimental and modeling study

of the diffusion path of Ce(III)-EDTA in compacted bentonite. Chem Geol 636:121639. https://doi.org/10.1016/j.chemgeo.2023.121639

41. Zhu T, Zhang Y, Tao C, Chen W, Cheng H (2023) Prediction of organic contaminant rejection by nanofiltration and reverse osmosis membranes using interpretable machine learning models. Sci Total Environ 857:159348. https://doi.org/10.1016/j.scitotenv.2022.159348

42. Van Loon LR, Glaus MA, Müller W (2007) Anion exclusion effects in compacted bentonites: towards a better understanding of anion diffusion. Appl Geochem 22(11):2536−2552. https://doi.org/10.1016/j.apgeochem.2007.07.008

43. Holmboe M, Wold S, Jonsson M (2012) Porosity investigation of compacted bentonite using XRD profile modeling. J Contam Hydrol 128(1−4):19−32. https://doi.org/10.1016/j.jconhyd.2011.10.005

44. Zheng L, Rutqvist J, Birkholzer JT, Liu HH (2015) On the impact of temperatures up to 200 °C in clay repositories with bentonite engineer barrier systems: A study with coupled thermal, hydrological, chemical, and mechanical modeling. Eng Geol 197:278−295. https://doi.org/10.1016/j.enggeo.2015.08.026